데이터 마이닝에서 IRG에 의한 효율적인 빈발항목 생성방법

허용도[†] · 이광형[††]

요 약

기존의 데이터 마이닝 방법들은 공통적으로 최소지지도(minimal support) 값의 변경에 의한 빈발항목 탐사의 비효율성, 불필요한 연관규칙의 생성으로 인한 불편성, 그리고 새로운 트랜잭션을 추가하게 되면 이전 탐사과정에서 발견한 결과를 재활용하기 어렵다는 문제점들을 가지고 있다. 본 연구에서는 이러한 문제점들을 해결할 수 있는 SPM-IRG 방법을 제안한다. SPM-IRG 방법은 최소지지도 값을 이용하지만 트랜잭션내의 각 항목에 대하여 다른 항목과의 직접적 · 간접적인 관련성을 파악한 후 빈발항목을 생성한다. 또한 관심 있는 항목에 대해서만 빈발항목을 구성할 수 있기 때문에 기존의 방법에서 발생하는 비효율성을 최소화할 수 있다.

A New Method for Efficiently Generating of Frequent Items by IRG in Data Mining

Her Young-Do[†] and Lee Kwang-Hyung[††]

ABSTRACT

The common problems found in the data mining methods current in use have following problems. First: It is ineffective in searching for frequent items due to changing of minimal support values. Second: It is not adaptable to occurring of unuseful relation rules. Third: It is very difficult to re-use preceding results while adding new transactions.

In this paper, we introduce a new method named as SPM-IRG(Selective Pattern Mining using Item Relation Graph), that is designed to solve above listed problems. SPM-IRG method creates a frequent items using minimal support values obtained by investigating direct or indirect relation of all items in transaction.

Moreover, the new method can minimize inefficiency of existing method by constructing frequent items using only the items that we are interested.

Key words: Data Mining, frequent items, relation rules, minimal support

1. 서 론

기업들이 경쟁력을 강화하기 위해서는 축적된 데이터를 분석하고 정보와 지식을 획득하는 능력을 보유해야 한다. 그러나 1990년대에는 데이터를 분석하여 정보와 지식을 획득하는 능력이 데이터를 획득하고 저장하는 능력에 훨씬 미달하는 '데이터 과잉문제(Data Glut Problem)'가 발생하였다[8]. 이러한 데이터 과잉문제는 방대한 양의 데이터에 내재된 정보와 지식을 발견하는 능력의 개선에 의해서 해결될 수 있는데, 데이터 마이닝(data mining)은 바로 이런 요구사항을 충족시키는 새로운 정보 기술의 활용방법이다. 데이터 마이닝은 대량의 실제 데이터로부터 이전에 잘 알려지지는 않았지만, 묵시적이고 잠재적으

[†] 건양대학교 IT학부 부교수
[††] (주)ECO 시스템개발실 부장

로 유용한 정보를 추출하는 작업이다.

최근 수년동안 다양한 분야에서 데이터 마이닝에 대한 연구가 활발하게 진행되어 오고 있다. 그 동안 제안된 다양한 데이터 마이닝 기법들은 탐사하고자 하는 빈발항목(frequent itemset)의 대상에 따라 빈발패턴 마이닝(Frequent Pattern Mining) 방법과 closed 패턴 마이닝(Closed Pattern Mining) 방법으로 분류된다. 빈발패턴 마이닝 방법들[2,3,7]은 모든 가능한 빈발항목을 생성하는 형태이고, Closed 패턴 마이닝 방법들[4,6]은 가능한 모든 빈발항목들중에서 closed 빈발항목만을 생성하는 형태이다[1]. 기존의 데이터 마이닝 방법들은 공통적으로 최소지지도(minimal support) 값의 변경에 의한 빈발항목 탐사의 비효율성, 불필요한 연관규칙의 생성으로 인한 불편성, 그리고 새로운 트랜잭션을 추가하게 되면 이전 탐사과정에서 발견한 결과를 재활용하기 어렵다는 문제점들을 가지고 있다.

따라서 본 연구의 2장에서는 데이터 마이닝의 기본 정의들과 기존 연구들의 문제점을 고찰하고, 3장에서는 이러한 문제점들을 해결할 수 있는 SPM-IRG(Selective Pattern Mining using Item Relation Graph) 방법을 제안한다. SPM-IRG 방법은 최소지지도 값을 이용하지만 트랜잭션내의 각 항목에 대하여 다른 항목과의 직접적·간접적인 관련성을 파악한 후 빈발항목을 생성한다. 마지막으로 4장에서는 제안한 SPM-IRG 방법의 특징 및 알고리즘에 대하여 설명한다.

2. 개념 정의 및 관련연구

2.1 개념 정의

$I=\{i_1, i_2, ..., i_n\}$를 항목(item)들의 집합, TDB(Transaction Database)를 n개의 트랜잭션들의 집합이라 하자. 각 트랜잭션 T는 I의 부분집합($T \subseteq I$)이고, 고유의 트랜잭션 번호(Tid)를 갖는다.

[정의 1] I의 멱집합(power set)의 부분집합 X를 항목집합(Itemset 또는 pattern)이라고 하고, 특별히 $|X|=k$인 X를 k-항목집합이라고 한다.

[정의 2] 항목집합 X가 $X \subseteq I$ 이고 임의의 트랜잭션 T에 대해서 $X \subseteq T$ 이면, 트랜잭션 T는 항목집합 X를 지지한다. 항목집합 X의 지지도는 TDB에서 X를 지지하는 트랜잭션들의 개수이며, sup[X]로 표기한다. 항목집합 X의 최소지지도는 사용자에 의해서 임의적으로 설정되는 값이며, sup-min으로 표기한다.

[정의 3] 항목집합 X의 지지도가 최소지지도보다 크면 즉, $sup[X] \geq min\text{-}sup$ 일 경우 항목집합 X를 빈발항목(frquent Itemset)이라 한다.

[정의 4] 항목집합 X, Y에 대해, 규칙은 "$R:X \rightarrow Y$" 형식의 함축이며, $X,Y \subseteq I$ 이고, $X \cap Y = \emptyset$이고 $Y \neq \emptyset$ 이다. 이때, X를 규칙의 조건부(antecedent), Y를 결과부(consequent)라고 한다.

[정의 5] TDB에 있는 규칙 $R:X \rightarrow Y$의 지지도는 $sup[X \cup Y]$로 정의한다. 규칙 $R:X \rightarrow Y$의 신뢰도는 conf[R]로 표기하며, X를 지지하는 트랜잭션 T에 대하여 Y를 지지할 조건부 확률로 정의한다. 즉, $conf[R]=p(Y \wedge X)/p(X)=sup[X \cup Y]/sup[X]$ 이다. 규칙 R의 신뢰도가 최소신뢰도보다 같거나 크면 즉, $conf[R] \geq conf\text{-}min$ 이면 규칙 R을 연관규칙이라 한다. 규칙 R의 최소 신뢰도는 사용자에 의해서 임의적으로 설정되는 값이며, conf-min으로 표기한다.

일반적으로, 연관규칙들은 빈발항목 탐색단계와 탐색된 빈발항목으로부터 연관규칙을 생성하는 두 단계 과정을 거쳐 얻어진다. 빈발항목 탐색단계에서는 미리 결정된 min-sup 값 이상의 트랜잭션 지지도를 갖는 항목집합들의 모든 부분 집합들이 빈발항목이 된다. 따라서 잠재적인 빈발항목의 수는 모든 항목들의 멱집합의 크기와 같다. 연관규칙 생성단계에서는 모든 빈발항목 집합 L에 대해서 L의 공집합이 아닌 부분집합 A를 찾는다. 탐색된 부분집합 A에 대하여, sup[A]에 대한 sup[L]의 비율이 적어도 최소신뢰도 이상이면($sup[L] / sup[A] \geq min\text{-}conf$), $A \rightarrow (L-A)$ 형태의 연관규칙을 생성한다.

[예] <표 1>의 TDB에 대해 min-sup=2, min-conf=50% 라고 할 때, 빈발항목 'acdf'에 대해 항목

표 1. 트랜잭션 데이터베이스 TDB

트랜잭션 번호	트랜잭션내의 항목들
10	a, f, d, c, e
20	b, e, a
30	c, f, e
40	a, d, c, f
50	c, e, f

'cf'의 경우 지지도는 4이고 빈발항목 'acdf'의 부분집합이다. 이 경우 규칙 "R:cf→ad"의 신뢰도는 sup[acdf]/sup[cf]=2/4≥50% 이므로 규칙 R은 연관규칙이 된다.

2.2 관련 연구

TDB에서 빈발항목을 생성하는 과정은 데이터 마이닝에서 핵심적인 기술로, 지금까지 다양한 형태의 빈발항목 탐사방법들이 제안, 연구되어 왔는데, 탐사결과로 생성되는 빈발항목의 범위에 따라 크게 빈발패턴 마이닝(frequent pattern mining)과 closed 패턴 마이닝(closed pattern mining)로 구분할 수 있다.

2.2.1 빈발패턴 마이닝

빈발패턴 마이닝은 트랜잭션내의 모든 빈발항목을 찾아내는 방법으로 후보집합을 이용하는 Apriori 방법과 FP-트리를 이용한 FP-growth 방법이 있다.

1) Apriori 방법 : Apriori 방법은 'level-wise 탐색' 형태의 반복적 탐사방법을 사용한다. 첫 단계에서 1-빈발항목 집합 L_1을 구성한다. L_2를 생성하기 위해서 L_1내의 모든 항목을 조인(join)하여 후보집합(C_1)을 생성한다. 이 후보집합에서 min-sup보다 낮은 항목과 불필요한 항목들을 제거하여 L_2를 생성한다. 그리고 같은 방법으로 L_2를 이용하여 L_3를 생성한다. 이러한 반복적인 수행을 더 이상의 빈발항목이 생성되지 않을 때까지 수행한다. 이 방법의 단점은 길이가 큰 항목집합을 탐사하기 위해서 반복적으로 TDB를 검사해야 하기 때문에 후보집합을 처리하는데 많은 오버헤드가 발생한다는 것이다.

2) FP-growth 방법 : FP-growth 방법은 prefix 트리 구조를 적용한 FP-트리(frequent pattern tree)를 이용한다. FP-트리의 노드는 1-빈발항목으로만 구성되며, 노드구성은 빈도수가 높은 노드가 낮은 노드보다 공유기회를 더 많이 갖도록 정렬된다. 즉, TDB를 검사하여 각 트랜잭션 내의 항목 중에서 min-sup 이하의 지지도를 갖는 항목들을 제거하고, 항목의 빈발횟수가 많은 순서로 트랜잭션들을 재구성하여 FP-트리를 구성한다. 빈발항목을 검사하기 위해서는 분할기반(partition-based)의 분할 후 정복(divide and conquer) 방법을 사용한다. 이 방법은 탐사범위인 조건부 패턴기저(conditional pattern base)의 범위를 줄여 준다. FP-growth 방법은 최소지지도에 의해 탐사시점에서 빈발항목들이 제거하기 때문에 최소지지도가 변경되거나 새로운 트랜잭션이 추가되면 전체 빈발항목 탐사과정을 다시 수행해야 한다는 단점이 있다.

2.2.2 Closed 패턴 마이닝

앞에서 설명한 빈발패턴 마이닝 방법은 매우 많은 수의 중복된 빈발항목을 생성하는 단점이 있다. Closed 패턴 마이닝은 중복성이 제거된 항목집합만을 생성하고, 보다 효율적으로 연관규칙을 생성할 수 있는 방법으로 A-Close, Charm, CLOSET 방법 등이 있다.

1) A-Close 방법 : 이 방법은 Apriori 방법의 변형형태로 Apriori 방법을 적용하여 후보집합을 생성하고 이를 이용하여 빈발항목들을 생성한다. 그리고 해당 항목을 포함하고 있는 모든 트랜잭션의 교집합을 수행하여 빈발 closed 항목집합을 생성한다. 이 방법은 여전히 Apriori 방법의 문제점을 포함한다.

2) Charm 방법 : 이 방법은 TDB를 수직데이타 형식(vertical data format)으로 변환하여 활용한다. 즉, 각 항목들은 트랜잭션 식별자(tid)들의 집합과 연관된다. Charm 방법은 첫 단계에서 1-빈발항목을 탐사하여 해당 항목을 가지(branch)로 갖는 트리를 구성한다. 그리고 다음 단계에서 항목집합의 구성을 위해 동일 레벨의 항목들을 조합한다. 그리고 이러한 과정을 각 가지에 대해 수행하여 전체 빈발 closed 항목집합을 선정한다. 이 방법의 문제점은 트랜잭션 식별자 집합(tid-set)의 교집합을 반복적으로 계산하면서 많은 오버헤드가 발생한다는 점이다.

3) CLOSET 방법 : 이 방법은 첫 단계에서 TDB를 검사하여 1-빈발항목의 집합 f-list(frequent item list)를 생성하는데, f-list의 구성은 지지도가 높은 항목 순으로 정렬된다. 그리고 다음 단계에서 f-list를 이용하여 탐사범위를 분할한 후 각 영역에 대하여 빈발 closed 항목집합의 부분집합을 찾는다. 빈발 closed 항목집합의 부분집합은 해당 조건부 패턴 기저(conditional pattern base)를 재귀적으로 구성함으로써 탐사할 수 있다.

표 2는 표 1의 TDB에 대해 각 방법의 결과로 얻어지는 빈발항목 생성결과를 나타낸다.

표 2. 각 방법의 빈발항목 생성 결과(표 1 이용)

방 법	생성된 전체 빈발항목(항목: 지지도)
Apriori	{a:3},{c:4},{d:4},{e:4},{f:4},{a,c:2},{a,d:2},{a,e:2},{a,f:2},{c,d:2},{c,e:3},{c,f:4},{d,f:2},{e,f:3},{a,c,d:2},{a,c,f:2},{a,d,f:2},{c,d,f:2},{c,e,f:3},{a,c,d,f:2}
FP-growth	{f,c,a,d:2},{c,a,d:2},{f,a,d:2},{a,d:2},{f,c,d:2},{c,d:2},{f,d: 2},{d:2},{f,c,a:2},{c,a:2},{e,a:2},{f,a:2},{a:3},{f,e,c:4},{e,c:3},{f,c:4},{c:4},{f,e:3},{e:4},{f:4}
A-Close	{a:3},{c,f:4},{a,c,d,f:2},{e:4},{a,e:2},{c,e,f:3}
Charm	{d,a,f,c},{a},{f,c},{a,e},{f,c,e},{e}
CLOSET	{c,f,a,d:2},{a:3},{e,a:2},{c,f:4},{c,e,f:3}, {e:4}

2.3 기존 방법들의 문제점

기존의 데이터 마이닝 알고리즘들은 기본적으로 다음과 같은 비효율성을 내재한다. 첫째 최소지지도가 변경되면 전체 TDB를 다시 탐사해야 하기 때문에 많은 오버헤드를 필요로 한다. 즉 기존의 방법은 빈발항목 생성단계에서 최소지지도 값보다 작은 항목집합들을 다음 단계에서 제거한다. 그러나 만약 최소지지도의 값이 이전에 설정된 값보다 작게 또는 크게 변경되면 이전 탐사에서 제외되었던 항목들이 탐사대상 범위에 포함되거나 또는 빈발항목에 포함되었던 것들이 제거되어야 하기 때문에 최소지지도 값의 변경으로 인해서 처음부터 다시 탐사과정을 수행해야 한다. 이러한 오버헤드는 재탐사 필요성을 제거함으로써 제거할 수 있다.

둘째, 새로운 트랜잭션의 추가에 의해 이전에 생성된 빈발항목을 재활용하지 못하고, 다시 처음부터 생성해야 한다. 왜냐하면, 빈발항목에서 제외되었던 항목들이 새로 추가되는 트랜잭션에 의해서 최소지지도 값보다 같거나 클 수 있기 때문이다. 따라서 한 단계의 결과를 다음 단계에서 활용하는 기존 알고리즘에서는 전체 탐사과정을 반드시 다시 수행해야 한다.

셋째, 사용자의 관심과 무관하게 데이터베이스내의 모든 항목을 대상으로 과도하게 생성되는 빈발항목 및 연관규칙으로 인해 관심이 있는 연관규칙을 추출하는데 어려움이 있다. 즉 데이터 마이닝 알고리즘에서 항목들간의 연관규칙을 생성하기 위해서는 트랜잭션내의 모든 항목들을 대상으로 반복적으로 최소지지도보다 큰 모든 빈발항목들을 생성하게 된다. 연관규칙은 이러한 빈발항목을 이용하여 생성하기 때문에 매우 많은 연관규칙이 생성되고 이로 인해 사용자가 관심이 있는 연관규칙만을 추출하는 것이 매우 어렵게 된다. 또한 탐사과정을 거쳐 생성된 빈발항목들은 경우에 따라서는 대부분이 사용자의 관심이 없는 것으로 구성될 수 있으며 이에 따라 과도하게 불필요한 연관규칙이 생성된다는 문제점이 있다.

3. IRG에 의한 새로운 빈발항목 생성방법

3.1 기본 개념

본 논문에서는 트랜잭션 내의 각 항목에 대하여 다른 항목과의 직·간접적인 관련성을 파악한 후 최소지지도의 값을 이용하여 빈발항목을 생성하는 SPM-IRG(Selective Pattern Mining using Item Relation Graph) 방법을 제안한다. SPM-IRG은 먼저 주어진 항목집합에 대해 IRG를 구성하고 단계적으로 확장하여 k-항목집합에 대한 지지도를 계산한 후, 최소지지도를 이용하여 빈발항목을 생성한다. 이 때, IRG는 항목간의 직접적인 관련성을 나타내며, 관련 있는 IRG들을 이용하여 항목간의 간접적인 관련성을 추출한다.

[정의 6] IRG(Item-Relation Graph)

IRG는 다음 두 종류의 노드로 구성된 그래프이다.

① 연관(association)노드 : 탐사 대상이 되는 항목으로 IRG의 중앙에 위치한다.

② 관련(related)노드 : 연관노드와 직접적으로 관련이 되는 항목으로, 연관노드와 동일한 트랜잭션에 존재하는 항목이다. 트랜잭션 내에서 연관노드와의 위치에 따라 좌-노드(left-side node, 트랜잭션 내에서 연관항목의 바로 이전에 위치하는 항목)와 우-노드(right-side node, 트랜잭션 내에서 연관항목의 바로 이후에 위치하는 항목)로 분류한다.

IRG를 구성하기 위해서는 먼저 각 트랜잭션 내의 항목들을 알파벳 순서로 정렬시킨다. 그 이유는 탐사하고자 하는 연관노드에 따라 탐사대상의 범위를 조절하기 위함이다.

IRG는 트랜잭션 내에서 연관노드와 관련노드의 발생빈도에 따라 가중치(weighted value : 노드간의 경로 수)를 할당한다. 가중치는 연관노드와 관련노드

의 위치에 따라 관련-연관노드 가중치(in-degree value)와 연관-관련노드 가중치(out-degree value)로 구분된다. 이 가중치는 항목집합의 지지도를 계산하는데 이용된다.

[예] <표 3>에서, 항목 c는 5개의 트랜잭션 중에서 4개의 트랜잭션에 존재하며, c의 전·후 항목을 이용하여 IRG를 구성하면 그림 1과 같다. 트랜잭션 번호 10, 40에 의해 관련-연관노드(a-c) 가중치는 2가 되고, 10, 30, 40, 50에 의해서 연관-관련노드(c-d, c-e) 가중치는 각각 2가 된다. 실선 링크는 연관노드의 좌-노드가 존재함을, 점선 링크는 좌-노드가 존재하지 않음을 나타낸다. 좌-노드와 우-노드의 가중치가 일치하지 않는 경우는 연관노드에 대하여 동일한 트랜잭션에 존재하지 않는 항목이 존재함을 나타낸다. 연관노드 d는 트랜잭션 번호 10, 40에 의해 관련-연관노드(c-d) 가중치는 2가 되고 연관-관련노드(d-e, d-f) 가중치는 각각 1이 된다. 또한 좌-노드와 우-노드의 가중치가 일치하기 때문에 모든 항목은 동일한 트랜잭션 내에 존재함을 나타낸다. 연관노드 d에 대한 IRG는 그림 2와 같으며, 전체 항목에 대한 IRG는 그림 3과 같다.

항목집합에 대한 지지도는 트랜잭션 내에서 관심대상의 항목들이 동시에 존재하는 개수를 의미한다. 즉, TDB에서 연관노드에서 관련노드로의 경로가 얼마나 존재하는 지를 의미한다. 크기가 1인 항목에 대한 지지도는 해당 IRG에서 직접적으로 구하는 반면 크기가 2이상의 항목집합에 대한 지지도는 노드의 확장 IRG를 이용하여 간접적으로 구할 수 있다.

표 3. 표 1의 TDB의 항목 정렬

트랜잭션 ID	트랜잭션 내의 항목들
10	a, c, d, e, f
20	a, b, e
30	c, e, f
40	a, c, d, f
50	c, e, f
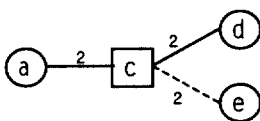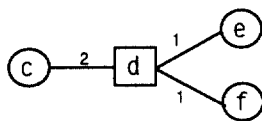
그림 1. c의 IRG 그림 2. d의 IRG
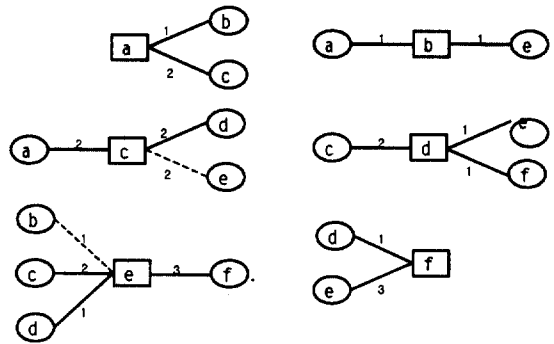
그림 3. 전체 IRG

1) 1-항목집합의 지지도 계산 : 관련-연관노드의 가중치의 합과 연관-관련노드의 가중치의 합 중 큰 값(max(sum of in-degree value, sum of out-degree value))을 지지도로 설정한다. 그림 1의 경우 연관노드 c에 대한 지지도는 4가 된다. 즉, 관련-연관노드(a-c:2)의 가중치는 2가 되고, 연관-관련노드(c-d:2, c-e:2)의 가중치는 4가 된다. 따라서 c의 지지도는 max(2, 4)=4가 된다.

2) k-항목집합의 지지도 계산(k≥2) : 탐사하고자 하는 항목집합을 $x_1x_2...x_k$ 라 할 때, 이 항목집합의 지지도는 다음과 같이 반복적으로 계산된다. x_1x_2의 지지도를 구한 후 $X_2x_3(X_2=x_1x_2)$의 지지도를 구하고, ..., $X_{k-1}x_k(X_{k-1}=x_{k-2}x_{k-1})$의 지지도를 구한다. 항목 쌍 x_ix_{i+1}의 지지도를 구할 때 x_{i+1}이 x_i의 IRG에 존재하는 경우(직접연결)와 x_{i+1}이 다른 노드 y_j의 IRG를 통해서 연결되는 경우(간접연결)가 존재한다. 직접연결의 경우 IRG에서 x_ix_{i+1}의 해당 가중치와 바로 전에 구한 $X_{i-1}x_i$까지의 지지도 중 적은 값을 지지도로 선택한다. 간접연결의 경우는 x_i에서 x_{i+1}로의 경로상의 모든 지지도를 더한 값을 x_ix_{i+1}의 지지도로 선택한다. 그리고 최종적으로 직접연결 지지도와 간접연결 지지도의 합을 더하여 k-항목집합의 지지도로 선택한다.

[예] 항목집합 'acdf'의 지지도 계산 : a의 IRG를 이용하여 관련노드 b와 c를 찾는다. b의 IRG에서 관련노드 e는 탐사항목 c보다 크기 때문에 (a-b-e)는 탐사과정에서 제외된다. 관련노드 c는 탐사항목 c와 동일하기 때문에 (a-c)의 확장 IRG를 구성하고 지지도를 계산한다(a-c:2). c의 IRG를 이용하여 관련노드 e를 찾는다. e는 탐사항목 d보다 크기 때문에 제외

하고, 같은 방법으로 (a-c-d)의 확장 IRG를 구성하고 min((a-c:2), (c-d:2))를 (a-c-d:2)의 지지도로 결정한다. d는 e와 f의 관련노드를 갖고 e는 탐사항목 f보다 작기 때문에 (a-c-d-e)의 확장 IRG를 구성하고 지지도를 계산한다. 즉, min((a-c-d:2), (d-e:1))= 1. 그리고 e는 탐사항목과 동일한 관련노드를 갖기 때문에 (a-c-d-e-f:1)의 확장 IRG를 구성한다. 또한 d의 다른 관련노드 f에 대해서 확장 IRG를 구성한다 (a-c-d-f:1). 결과적으로 'acdf'의 지지도는 (a-c-d-e-f:1) + (a-c-d-f:1)=2가 된다.
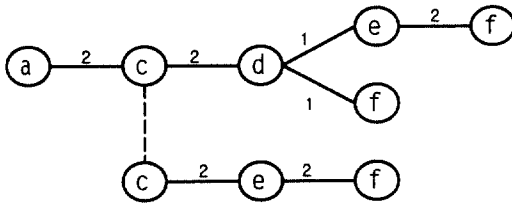
그림 4. acdf의 확장 IRG

4. SPM-IRG 알고리즘의 표현 및 특징

3장에서 설명한 SPM-IRG 방법의 알고리즘 표현은 다음과 같다.

```
[SPM-IRG 방법의 알고리즘 표현]
Procedure SPM_IRG(Itemset S){//S=x_1x_2...x_n
for(i = 1; x_i < x_{i+1}; i++){//x_i와x_{i+1}의 지지도 계산
x_i의 관련노드(우-노드)를 r_j(j≥2)라 하면
for(j = 1; r_j < x_{i+1}; j++) {
      case (r_j < x_{i+1}) //x_i와 x_{i+1}의 간접연결 형태
          r_j의 확장 IRG 구성;
          support_j = r_j의 가중치;
          indirect-support(r_j) 호출;
      case (r_j = x_{i+1}) // x_i와 x_{i+1}의 직접연결 형태
             x_i와 x_{i+1}의 확장 IRG 구성;
             support_i = x_i와 x_{i+1}의 가중치;
   }
  support = support_i + support_j
  if(support < min-sup)
     output 비-빈발항목
  }
 }
 Procedure indirect-support(IS){//IS의 관련노드
                                          // =s_1s_2...s_m
  for (k = 1; s_k < x_{i+1}; I++){
    case (s_k < x_{i+1})
         r_k의 확장 IRG 구성;
         support_k = s_k의 가중치;
         support_j = min(support_k, support_j);
         indirect-support(s_k) 호출;
    case (s_k = x_{i+1})
         s_k의 확장 IRG 구성;
         support_k = min(support_k, support_j);
    case (s_k > x_{i+1})
         exit    // 비-빈발항목
  }
 }
```

표 1의 TDB를 이용하여 관심 있는 항목 acdf에 대해 SPM-IRG 알고리즘을 수행하여 얻은 빈발항목의 결과는 다음과 같다.

{a : 3}, {c : 4}, {d : 4}, {f : 4}
{a,c : 2}, {a,d : 2}, {a,f : 2}, {c,d : 2},
{c,f : 4}, {d,f : 2}
{a,c,d : 2}, {a,c,f : 2},{a,d,f : 2}, {c,d,f : 2}
{a,c,d,f : 2}

SPM-IRG 기법은 사용자의 관심대상이 되는 항목에 대해서만 탐사를 수행하기 때문에 항목 'e'에 대해서는 탐사를 수행하지 않는다. 만약에 관심대상이 'abcdef'로 변경되거나 최소지지도가 변경되어도 이전 탐사 결과를 활용하여 빠르게 탐사 결과를 얻을 수 있다. 본 논문에서 제안한 SPM-IRG 방법은 기존 방법들과 비교할 때 다음과 같은 특성을 갖는다.

첫째, 탐사하고자 하는 항목과 그와 관련된 항목에 대해서만 빈발항목을 생성하기 때문에 사용자의 관심이 있는 연관규칙만을 생성할 수 있다. 따라서 모든 빈발항목을 생성하고 그에 대한 모든 연관규칙을 생성하는 알고리즘에 비해 효율적으로 관심이 있는 연관규칙만을 추출할 수 있다. 또한 선택된 탐사항목에 따라 전체 탐사범위를 효과적으로 줄일 수 있다. 즉, 선택된 항목들의 사전식 순서가 낮을수록 탐사범위는 축소된다. 예를 들면 항목집합 I = {a,b, ... , z}에 대해 탐사하고자 하는 항목이 {a,b,c}라면 탐사과정에서 {d,e,f, ... , z}의 항목들에 대해서는 고려할 필요가 없다.

둘째, 새로운 트랜잭션이 추가되더라도 이전에 생성된 탐사과정을 재활용할 수 있기 때문에 전체 탐사과정을 처음부터 수행해야 하는 기존 알고리즘의 오버헤드를 획기적으로 줄일 수 있다. 예를 들면 표 2의 TDB에서 'adfg'와 'adcef'가 추가되고, 'acdef'의 지지도를 계산한다면, 'acdf'의 확장 IRG에 새로운 트랜잭션의 항목들을 추가해서 지지도를 계산하면 된다..
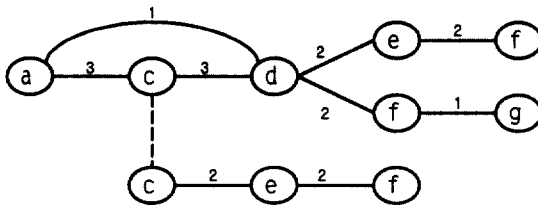
그림 5. acdef의 확장 IRG

셋째, 최소지지도의 값이 변경되더라도 단지 확장 IRG상의 가중치와 최소지지도와의 비교를 통해서 빈발항목을 재 계산할 수 있다. 기존의 알고리즘에서 빈발항목을 구하는 방식은 작은 크기의 빈발항목을 구하고 이를 기반으로 단계적으로 큰 크기의 빈발항목을 생성하게 된다. 즉, 한 단계에서 최소지지도 보다 작은 항목집합들은 다음 단계에서 제거된다. 따라서 만약 최소지지도 값이 이전 값보다 작게되면 이전 탐사 기준에서 제외되었던 항목집합들이 탐사 대상 범위에 포함되어야 하기 때문에 처음부터 다시 탐사과정을 거쳐야 한다. 그러나 SPM-IRG 방법은 기존의 방법들처럼 단계마다 생성된 항목집합을 최소지지도에 따라 빈발항목을 구성하는 것이 아니라 먼저 탐사하고자 하는 항목간의 각각의 지지도를 구한 후 이를 더하여 최소지지도와 비교하기 때문에 최소지지도 값의 변경에 따라 다시 탐사할 필요가 없다.

5. 결론 및 향후 연구방향

본 연구에서는 기존의 마이닝 기법에서 발생하는 최소지지도 값의 변경에 의한 비효율성, 불필요한 연관규칙의 생성으로 인한 불편성, 이전에 생성된 빈발항목의 재활용의 어려움과 같은 문제점을 해결하기 위한 방법으로 SPM-IRG 방법을 제시하였다.

이 기법은 탐사 시 전체항목을 대상으로 하지 않고 관심 대상이 되는 항목에 대해서만 탐사를 수행한다. 즉, 관심항목에 따라 탐사 범위가 좌우된다. 따라서 불필요한 연관규칙의 생성을 줄일 수 있다. 또한 기존 방법과는 달리 비-빈발항목을 탐사과정에서 제거하지 않기 때문에 최소지지도가 변경되거나 새로운 항목이 추가 또는 기존의 항목이 삭제되는 경우에도 이전 탐사 결과를 재활용할 수 있다.

그러나 탐사결과를 얻는데 있어서, 중복되는 부분의 제거를 고려하지 않았기 때문에 관심대상의 항목들이 중복되어 나타나는 경우가 발생한다. 따라서 Closed 빈발항목만을 생성하기 위한 방법을 앞으로 고려해야 한다.

참 고 문 헌

[1] R. Mao, "Adaptive -EP : An Efficient and Effective Method for Multi-level Multi-dimensional Frequent Pattern Mining", MS thesis, Simon Franster Univ. April, 2001

[2] R.Agrawal, S. Srikant, "Fast Algorithms for Mining Association Rules", Proc. 1994 Int. Conf. on Very Large Data Base, pp. 487-499, Santiago, Chile, Sept. 1994

[3] J. Han, J. Pei, Y. Yin, "Mining Frequent Patterns without Candidate Generation", Proc. 2000, ACM-SIGMOD Int. Conf. on Management of Data(SIGMOD'2000), pp. 1-12, Dallas TX May 2000

[4] N. Pasquier, Y. Bastide, R. Taouil, L. Lakhal, "Discovering Frequent Closed Itemsets for Association Rules", proc. 7th Int. Conf. on Database Theory(LCDT'99), pp.398-416, Jerusalem, Israel, Jan 1999

[5] M. Zaki, C. Hssial, "CHARM : An Efficient Algorithm for Closed Association Rule Mining", Tech. Rpt. 99-10, Computer Science, Rensselaer Poly-Technic Institute 1999

[6] J. Pei, J. Han, R.Mao, "CLOSET : An Efficient Algorithm for Mining Frequent Closed Itemset", Proc. 2000 ACM-SIGMOD, Int. Workshop on Data Mining and Knowledge Discovery(DUKD'00), Dallas, TX, May 20001.

[7] H. Toivonen, "Sampling Large Databases for Association Rules", Proc. 1996 Int. Conf. on Very Large Data Base(VLDB'96), pp.134-145, Bombay, India, Sept. 1996

[8] Fayyad, U. M., G. Piatetsky-Shapiro, and P.Smyth, "from Data Mining to Knowledge Discovery", In Advances in Knowledge discovery and Data Mining, AAAI Press/MIT Press, CA., pp.1-34, 1996

허 용 도

1986년 고려대학교 수학과(이학사)
1988년 고려대학교 수학과 전산학전공(이학석사)
1993년 고려대학교 수학과 전산학전공(이학박사)
1992년~현재 건양대학교 IT학부 부교수
관심분야 : 분산시스템, 컴퓨터 네트워크 보안, 데이터마이닝, 정보검색

이 광 형

1988년 순천향대학교 전산학과(공학사)
1991년 고려대학교 전산과학과(이학석사)
1995년 고려대학교 전산과학과(이학박사)
1996년~1997년 한국산업표준원 선임연구원
1997년~2000년 동서대학교 컴퓨터공학과 전임강사
2001년~현재 (주)ECO 시스템개발실 부장
관심분야 : 분산시스템, 컴포넌트 기반 개발 방법론, 데이터마이닝, 정보검색

논 문 정 정 안 내

2001년 12월에 발간된 멀티미디어학회논문지 제4권 제6호에 게재된 "경계 주사 구조를 이용한 새로운 실시간 모니터링 실장 제어기 설계" 논문에서 다음과 같은 연구비 지원 내용이 편집 및 교정 과정에서 누락되었음을 알려드립니다.

■ 누락 문구

이 논문은 1999년도 안동대학교 학술연구조성비 지원에 의해서 연구되었음.